Homework 1 Report - PM2.5 Prediction

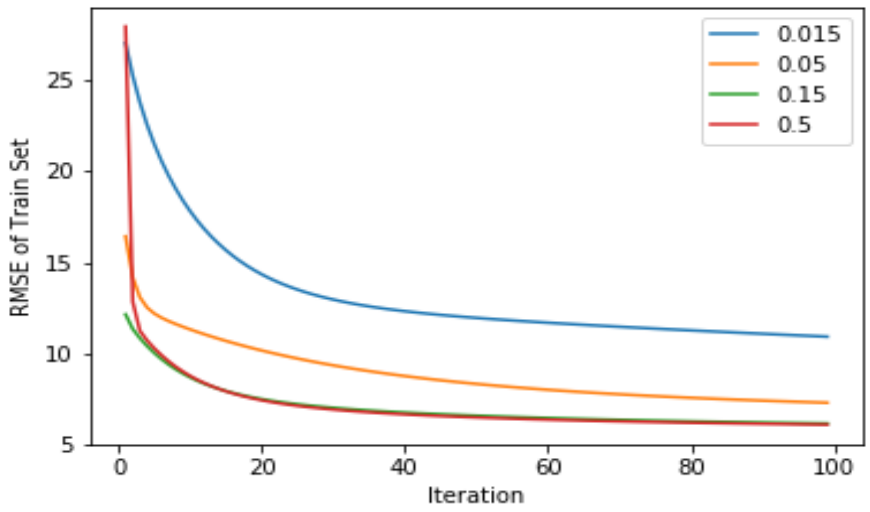
學號：r06323011 系級：經濟碩一 姓名：葉政維

1. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

本題經題（4）所述相同之 preprocessing，使用原資料中的 18 個 features 以及常數項，但其中的風力、風向被轉換為 x 與 y 方向之分量（含小時版本，共 4 項）。皆使用 linear regression （lambda = 0）。此外皆使用 adaGD 並在最終模型跑 10 萬次迴圈。

Full Model 中，public and private score 分別為 6.00927 與 6.33144；PM2.5 only 的版本中兩個 score 則為 7.21057 與 6.82601。Full Model 的表現皆較 PM2.5 好，即使不使用 regularization。這也許是因為資料量夠大，相較之下，163 個 features 還不至於導致 Overfitting。值得一提的是，如果 adaGD 得迴圈數太少，Full Model 似乎會被低估。

2. (2%) 請分別使用至少四種不同數值的 learning rate 進行 training（其他參數需一致），作圖並且討論其收斂過程。



上圖使用 adaGradient Descend 方法並搭配不同的 learning rate，在每個 iteration 下繪製 Training Set 的 RMSE。當 learning rate 為 0.015 ，顯然步伐太小了，而其他較大（甚至大很多）的 learning rate 最終都會收斂。倘若不使用 adaGradient Descend 且沒有 time decreasing 的 learning rate，其實非常容易使用到不會收斂的 learning rate。

3. (1%) 請分別使用至少四種不同數值的 regulization parameter λ 進行 training（其他參數需一至），討論其 root mean-square error（根據 kaggle 上的 public/private score）。

使用題（1）中的 Full Mode，並使用 lambda 為 0、100、1000、10000，最終模型皆在 adaGD 之下跑 10 萬個迴圈。對應之 Private Score 依序為 6.09927、5.98953、6.03968、6.20078，對應之 Public Score 為 6.33144、6.333545、6.33797、6.35247。略有差異，但至少顯示過小的 lambda 不一定最好（當 adaGD 迴圈夠多時，這點會比較明顯），但過大的 lambda 則不會有好結果。概念上，過大的 lambda 很可能將有一定預測能力的變數過度往零靠攏，預測能力因此下降。

4. (1%) 請這次作業你的 best_hw1.sh 是如何實作的？（e.g. 有無對 Data 做任何 Preprocessing？Features 的選用有無任何考量？訓練相關參數的選用有無任何依據？）

首先，先對資料進行清理：將不合理的負值調整為零；對某些變數進行必要之轉換如（風向為 0~360 不連續的值，將其轉為弧度 radian 以及 xy 軸上的分量）。此外，資料中有許多變數值集中於零，而且靠近零的測量值卻相當稀少，甚至有些樣本點有多數的值皆為零。**我認為這是 missing value 呈現到我們面前時被誤 code 成 0**，具體處理方式是將值為零的樣本點以前一筆有效樣本去推估，我對降雨量之外的變數都進行如此處理。從 Validating error 來看，如此帶來不少的改善，儘管因為 PM2.5 的值被更動，Validating error 恐有低估之虞。（我對 testing data 也進行同樣的處理，具體做法是在每筆 id 之內向前補，若為 id 的第一筆（lag9）則向後補）

其次，依照 test data 的結構，逐月對每個變數生成 lag 9 期的變數。在挑選模型上，我從全部 lag 9 期的變數開始，逐步減少 lag 其數。最後依照 validating error 判斷，僅僅 lag1 就有不錯的表現（註：事實上，當初在做此篩選時，並沒有做好 Preprocessing，事後再做一次發現表現最好的並非 lag1 期，此外我猜測：adaGD 的迴圈太少有可能導致一些複雜的模型被低估）。接著，我在 lag1 模型的基礎上，嘗試是否加入變數會有更好結果，如 lag 1 期的二次項、9 期平均、lag 2 期。事實上，我沒有一個好的流程去探索每一種潛在的可能，僅是在前次更新的結果上，貪婪式地想辦法再做得更好。Validating error 也顯示，更動 features 並未如 preprocessing data 帶來如此大的改善。

有關 feature 挑選的一些直覺，我認為風、雨可能是影響 PM2.5 相當大的因素。逐月檢查平均 PM2.5 時，可發現雨季的 PM2.5 都相當低，因此在嘗試生成的變數中，有許多圍繞著 RAINFALL 與風力，如兩者的交乘項。儘管從 Validating error 看來，添加這些變數並沒有很大的改善，也許是因為現有的變數已經足以捕捉到我所想的關係了。

最後有關訓練參數，我使用 Ridge Regression 以及給定一些候選的參數 lambda 值，並以 10 fold CV 挑選 validating error 最小的參數，最後再以完整的 training data 得出最終係數 hypothesis。